Identification in silico du site de fixation à l'ARN de la protéine PTBP1 : découverte de motifs par utilisation de l'outil RSAT via des données CLIP-Seq publiques

Léa Fléchon, Stéphanie Mottier, Aymeric Antoine-Lorquin, Catherine Belleannée

Master 1 Bio-informatique et Génomique

Université de Rennes 1

Année universitaire 2013-2014

Identification in silico du site de fixation à l'ARN de la protéine PTBP1 : découverte de motifs par utilisation de l'outil RSAT via des données CLIP-Seq publiques

Mémoire présenté par :

Léa FLÉCHON

Encadrement scientifique :

Catherine BELLEANNÉE (Équipe DYLISS, INRIA, Rennes)

Stéphanie MOTTIER (Équipe EGD, Institut Génétique et Développement, Rennes)

UMR CNRS *6074* Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)

UMR CNRS 6290
Institut de Génétique et Développement de Rennes (IGDR)

Annabelle MONNIER
Equipe Génomique Fonctionnelle Intégrée et Biomarqueurs

UNIVERSITÉ DE RENNES 1

ENGAGEMENT DE NON PLAGIAT

Je, soussigné(e) ……..... étudiant(e) en…….. déclare être pleinement informé que le plagiat de documents ou d'une partie de document publiés sur toute forme de support, y comprit l'internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée.

En conséquence, je m'engage à citer toutes les sources que j'ai utilisées pour la rédaction de ce document.

Date :

Signature :

Document à compléter de manière manuscrite et à insérer obligatoirement en première page du rapport de stage.

Laboratoire de Génomique Médicale
BMT-HC - CHU Pontchaillou

2 rue Henri le Guilloux
35033 Rennes Cedex
FRANCE

Annabelle MONNIER
annabelle.monnier@univ-rennes1.fr
TÉL. 33 (0)2 99 28 92 54

Remerciements

Je souhaite tout d’abord remercier mes maîtres de stage, Catherine Belleannée et Stéphanie Mottier. Leur présence et leur soutien constant m’ont été d’une grande aide dans la rédaction de ce rapport. Merci également pour leur disponibilité, leur attention et leur suivi continu de mon sujet. Un sincère remerciement tout particulièrement à Catherine pour son investissement, son attention et ses encouragements. Je remercie aussi en particulier Stéphanie pour l’ensemble des documents envoyés et pour la relecture minutieuse de ce rapport. Remerciement également à Aymeric Antoine-Lorquin pour sa gentillesse, sa bonne humeur, son aide pour la manipulation informatique des données.

Je remercie Serge Hardy pour avoir suivi avec intérêt le stage depuis le tout début et pour son aide dans l’amélioration des résultats.

Un grand merci à Jacques van Helden pour sa disponibilité, ses réponses rapides par mail et ses judicieux conseils qui m’ont permis d’interpréter de nombreuses données anormales.

Merci à toute l’équipe Dyliss pour l’accueil et la bonne ambiance générale. Et merci aux étudiants de M2, fort bien sympathiques de la pièce d’à côté, et de M1 pour les discussions et leur accessibilité, ce qui m’a permis de réaliser mon stage dans les meilleures conditions possibles !

Sommaire

Introduction

1) La protéine de liaison à l'ARN PTBP1

La **PTBP1** (Polypyrimidine Tract-Binding Protein 1) est une protéine de liaison à l'ARN (RNA Binding Protein, RBP) qui participe à la régulation post-transcriptionnelle de l'expression des gènes. Elle a été initialement caractérisée par sa capacité à interagir avec la région riche en pyrimidines située en amont du site 3' d'épissage d'un pré-ARNm (Garcia-Blanco *et al.* 1989). Par la suite, d'autres études ont montré son rôle régulateur dans de nombreux domaines : contrôle de l'épissage alternatif des exons (Xue *et al.* 2009), polyadénylation, transport et stabilité des ARNm, traduction contrôlée par des ARNs structuraux IRES (Site d'Entrée Interne des Ribosomes) (Glisovic *et al.* 2008).

La PTBP1 possède quatre motifs de liaison à l'ARN de type **RRMs** (RNA Recognition Motifs, **Figure 1**) qui vont lui permettre de se fixer sur des séquences riches en CU afin de réguler l'utilisation de sites d'épissage avoisinants. Chaque RRMs est constitué de 80 à 90 acides aminés dont la structure a permis de préciser leur interaction avec l'ARN (**Figure 1**).

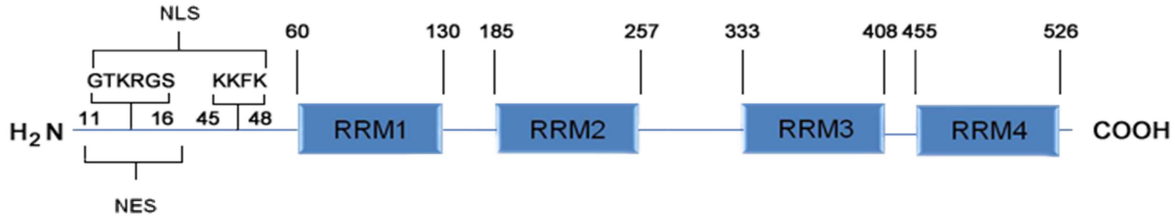


Figure 1 : Représentation schématique de la structure de PTBP1 (Sawicka *et al.* 2008).

Il existe des espaces entre les RRM1 et 2, et RRM2 et 3. Les RMM3 et 4 sont juxtaposés et possèdent des structures hydrophobes qui permettent la courbure de l'ARN sur lequel ils se fixent.

2) Caractéristiques du site de liaison à la PTBP1

a) Expérience CLIP-Seq : localisation du site de fixation

En 2009, une première étude à haut débit reposant sur la technique du **CLIP-Seq** (CrossLinking and ImmunoPrecipitation) a permis d'identifier les régions ciblées et régulées par PTBP1 dans les cellules HeLa (Xue *et al.* 2009). Cette méthode a été développée dans l'objectif d'identifier exhaustivement ***in vivo*** l'ensemble des sites de contacts entre une protéine de liaison à l'ARN et ses cibles ARN (Ule *et al.* 2005). Le CLIP débute par la formation de liaisons covalentes entre les ribonucléotides de l'ARN et les acides aminés des

protéines en exposant la cellule à des UV de type C (254 nm). Les complexes ARN/protéines sont alors immunoprécipités à l'aide d'anti-corps de la protéine d'intérêt. Après action de la protéinaseK, les ARNs sont ensuite rétrotranscrits pour être séquencés. Les données brutes du CLIP-Seq obtenues dans cette étude sont un grand nombre de petits fragments d'ADNc d'une taille entre 30 et 50 nucléotides.

L'étude a montré que la protéine se lie à près de 50% des transcrits codant des protéines chez l'homme et qu'environ 30% des sites de liaison sont retrouvés dans des régions soumises à l'épissage alternatif, soulignant l'effet majeur de cette protéine dans la régulation de l'épissage. Elle a également montré que les sites de liaison à la PTBP1 sont le plus souvent présents en plusieurs copies dans les introns qui jouxtent l'exon régulé et dans une moindre proportion dans l'exon alternatif (Xue *et al.* 2009).

b) Analyse structurale du site de fixation

Les analyses structurales du site de liaison à la PTBP1 ont montré qu'il devait s'étendre au minimum sur une trentaine de nucléotides (Amir-Ahmady *et al.* 2005). De plus, la distance minimale entre les RRM3 et 4 est approximativement de 15 nucléotides (Oberstrass *et al.* 2005).

c) Motifs consensuels déjà identifiés

Plusieurs études réalisées *in vitro* et *in vivo* ont permis de préciser des motifs consensuels **riches en pyrimidine** c'est à dire avec des successions de C et U pour la liaison de PTBP1 (**Figure 2**). Plus récemment, il a également été montré que PTBP1 peut se lier à des motifs présentant **quelques insertions de résidus G** (Han *et al.* 2014).

Motifs connus pour être fixés par PTBP1		
Séquence	**Méthode**	**Publication**
UYUYU	CLIP-Seq	(Xue *et al.* 2009),
CUCUCUn{15}CUCUCU	SELEX (Systematic Evolution of Ligands by developmental enrichment)	(Oberstrass *et al.* 2005)

Figure 2 : Table des séquences consensus établies pour le site de fixation de PTBP1. Un Y correspond à une pyrimidine. « n{15} » signifie 15 nucléotides aléatoires.

Cependant ces motifs consensuels ne se sont pas révélés suffisants ou puissants pour prédire bio-informatiquement les sites réellement utilisés *in vivo.*

3) Objectifs du stage

Bien que l'arrivée des nouvelles technologies de séquençage à haut débit ait permis d'augmenter les connaissances sur le rôle de PTBP1 dans un type cellulaire donné, beaucoup de progrès reste à faire en ce qui concerne la prédiction bio-informatique de son rôle dans un autre type cellulaire ou chez une autre espèce.

Durant ce stage, nous avons souhaité mettre au point une nouvelle méthodologie bio-informatique d'identification de sites fonctionnels de PTBP1 à partir des données de CLIP-Seq de la protéine (Xue *et al.* 2009). La recherche des sites de liaison fonctionnels peut être appréhendée par l'utilisation d'outils informatiques dits de découverte de motif (**pattern discovery** en anglais). Cette technique permet de détecter les motifs représentatifs (fréquents) d'un jeu de données. Dans cette étude, les analyses ont été réalisées par l'intermédiaire d'une suite logicielle très complète **RSAT** (Regulatory Sequence Analysis Tools, cf Méthode) (van Helden 2003) qui est actuellement centrée sur l'identification des sites de liaison des protéines à l'ADN à partir de données CHIP-Seq.

Plus précisément, le stage vise deux objectifs complémentaires.

Le premier objectif est biologique. Il s'agit de réexploiter les données publiques de CLIP-Seq de PTBP1 en utilisant de méthodes d'analyses récentes afin d'affiner la connaissance des motifs de fixation PTBP1. La finalité étant d'aboutir à la prédiction *in silico* de la fixation de la PTBP1 sur des génomes entiers annotés. Une partie critique de ce travail repose sur la constitution de jeux de données pertinents. Ainsi, de nouveaux jeux de données positifs ont été produits, par réassemblage des données brutes publiées (les reads). Par ailleurs, plusieurs jeux de données négatifs ont également été générés car ils sont déterminants pour s'assurer que les motifs prédits soient bien représentatifs des données positives, et qu'ils permettent ainsi la prédiction *in silico* de la fixation de la PTBP1.

Le deuxième objectif du stage est méthodologique. Il s'agit de mettre au point une méthodologie adaptée aux données de CLIP-Seq dans la suite logicielle RSAT en collaboration avec l'auteur de cet outil. En particulier, il s'agit d'explorer l'outil «**Peak-motifs**» de RSAT, dédié actuellement à la découverte de motifs sur des données CHIP-Seq (fixation de protéines à l'ADN). Dans le cas du CHIP-Seq, il s'agit de rechercher des sites de fixation principalement dans les zones non codantes alors que dans le cas du CLIP-Seq, les sites de fixation sont recherchés principalement dans les parties transcrites des génomes. De telles spécificités peuvent amener à adapter les procédures (paramétrage des outils, types de jeux de données de contrôle). Une démarche exploratoire sur les données PTBP1 permet de faire cette étude d'impact.

Matériels et Méthodes.

Matériels (données publiques accessibles via Gene Expression Omnibus : accession number GSE19323)

1) Jeu de données positif initial : « Pics_publi »

Le jeu de séquences positif initial a été réalisé à partir de deux expériences de CLIP-Seq en tant que monomères et deux expériences de CLIP-Seq en tant que dimères de PTBP1 (car la protéine a la capacité de se dimériser). Il en résulte l'identification de **51 394 pics fusionnés** qui sont des régions contenant une haute densité de reads assemblés sur le génome humain de référence (hg18) (Xue *et al.* 2009). Ces données sont contenues dans un unique fichier au format .bed qui donne les coordonnées chromosomiques et le sens des pics fusionnés.

2) Jeux de données positifs générés à partir des données brutes de séquençage

Le but de cette étude étant de maîtriser le pipeline complet d'analyse de données de CLIP-Seq afin d'en extraire les motifs le plus discriminant possible, nous sommes repartis de données brutes de séquençage (les reads) afin de générer nos propres jeux de pics fusionnés. Cela nous a permis d'utiliser des outils développés récemment pour les analyses de séquençage à haut débit (Galaxy : http://galaxy.nbic.nl/) mais aussi de faire une analyse plus fine des données en séparant les données de CLIP-Seq de PTBP1 en tant que monomère ou dimère.

Lors de cette réanalyse, les reads ont été filtrés par rapport à leur qualité moyenne > 20 (outil Sickle), puis les adaptateurs de séquençage aux extrémités 3’ et 5’ des reads ainsi que les séquences polyC ont été coupés (outil Cutadapt). Les reads ont alors été alignés sur le génome humain de référence hg19 (outil Tophat). La position des reads sur le génome a alors permis de déterminer les régions géniques des pics grâce à un logiciel de **peak-calling** (outil MACS : *http://liulab.dfci.harvard.edu/MACS/*). Trois nouveaux jeux positifs ont ainsi été générés.

« pics_new_monomères » : Jeu positif réalisé à partir des expériences de CLIP-Seq pour les protéines PTBP1 monomères. Il en résulte l'identification de **2339 pics.**

« pics_new_dimères » : Jeu positif réalisé à partir des expériences de CLIP-Seq pour les protéines PTBP1 dimères. Il en résulte l'identification de **298 pics.**

« pics_new_all » : Jeu qui fusionne les deux jeux précédents. Il contient au total **2637 pics**.

3) Jeux de données négatifs

Afin de discriminer avec précision les motifs présents uniquement dans nos jeux positifs, nous avons besoin de mettre en place des jeux négatifs de contrôle contre lesquels tester nos séquences positives. Dans cet objectif, quatre jeux négatifs ont été construits.

a) Jeux négatifs de séquences artificielles

« Jeu_négatif_shuffleseq » : Shuffle des données positives (outil **Shuffleseq de la suite** EMBOSS : http://emboss.bioinformatics.nl/cgi-bin/emboss/shuffleseq). La redistribution aléatoire des nucléotides permet de vérifier qu'un motif rencontré n'est pas dû au hasard de la distribution intrinsèque des nucléotides.

« Jeu_négatif_randomseq » : Séquences artificielles mimant la composition nucléotidique de l'organisme de référence (outil **Random_sequence** de la suite RSAT : http://rsat.ulb.ac.be/random-seq_form.cgi, option « Organism-specific Markov model »). Ces séquences sont obtenues par un processus de chaîne de Markov où les probabilités des nucléotides varient à chaque position. Par exemple, un ordre de Markov égal à 5 signifie qu'à chaque position le nucléotide va dépendre de la position des 5 nucléotides précédents. L'utilisation de ce jeu négatif a pour but de vérifier si les outils de RSAT permettant la fabrication de jeux négatifs sont adaptés aux données CLIP-Seq. Avec un ordre de Markov égal à 5, le jeu sera différent d'un simple Shuffle et commencera à mimer des séquences biologiques même s'il reste aléatoire.

b) Jeux négatifs issus de séquences réelles

« Jeu_négatif_Random_genome_fragments » : Fragments génomiques réels ayant la même distribution de taille que le jeu positif (outil **Random_genome_fragments** de la suite RSAT : http://rsat.ulb.ac.be/random-genome-fragments_form.cgi). L'outil va piocher au hasard ces fragments n'importe où dans un génome donné (ici hg19) sans contrôle possible des régions ciblées. Ce jeu, composé de vraies séquences biologiques quelconques, est utilisé dans les analyses de CHIP-Seq. Nous voulions tester la pertinence d'un tel jeu sur du CLIP-Seq.

« Jeu_négatif_Hela_genic_fragments » : Fragments géniques réels ayant la même distribution de taille que le jeu positif. Les fragments sont issus des séquences géniques de gènes exprimés dans Hela et non représentés dans le jeu positif de pics (outil **BEDTools** : http://bedtools.readthedocs.org/ en/latest/content/tools/intersect.html). BEDTools permet

d’associer les coordonnées contenues dans un fichier bed (ici le fichier positif des pics) avec des annotations d’un génome contenue dans un fichier GFF (ici génome humain d'UCSC *Genome Browser* : http://genome.ucsc.edu/). On obtient ainsi la liste des gènes représentés dans le jeu positif. On établit alors la liste complémentaire, contenant les gènes exprimés dans les cellules Hela mais négatifs vis-à-vis de PTBP1, convertie en fichier fasta sous Biomart (http://www.ensembl.org/biomart/martview/). À partir des séquences de « ces gènes négatifs», le jeu négatif est alors constitué en y prélevant aléatoirement des fragments d'une distribution de taille similaire à celle des pics fusionnés. Ce jeu a été imaginé pour se conformer aux données CLIP-Seq, qui concernent des données transcrites.

Méthodes : suite logicielle RSAT (*Regulatory Sequence Analysis Tools)*

La suite logicielle RSAT (van Helden 2003) a été développée pour rechercher des motifs surreprésentés dans un jeu de séquences en employant des méthodes statistiques fiables. Bien que RSAT ait été originellement développé pour l’analyse CHIP-Seq, son concepteur, avec qui nous collaborons pour ce stage, vise à la rendre utilisable pour des analyses de CLIP-Seq.

L’outil « **Peak-motifs** » de RSAT est un pipeline utilisé pour la découverte de motifs dans un ensemble de séquences. Il combine de nombreuses approches puissantes pour extraire les motifs surreprésentés dans le jeu de séquences. « **Peak-motifs** » fournit pour un jeu de données : le meilleur motif trouvé, sa séquence consensus, les 3 meilleurs mots/dyads avec leur index de significativité (Sig) qui est positif quand un mot est surreprésenté (exemple : Sig > 10 signifie un 1 faux-positif toutes les 10^{10} analyses). Un Sig très significatif apparaît en gras et rouge quand il est très significatif dans le tableau de résultat (= valeur entre 75 et 350). Peak-motif teste également chaque motif trouvé contre la base de données de facteurs de transcription (FT) « JASPAR Core Vertebrate » par un test de corrélation de Pearson (le match est considéré comme significatif à partir d’une valeur de 0,9), l’objectif étant d'identifier des liens potentiels avec des facteurs déjà connus (comme les FT). Les pics d’entrées et les sites prédits sont visualisables dans leur contexte biologiques par l'*UCSC Genome Browser* (Fujita & *al.* 2010). Trois des différents algorithmes sur lesquels Peak-motifs s'appuie sont présentés ci-dessous.

1) **Oligo-analysis** (van Helden *et al.* 1998)

Il s'agit d'une méthode rapide et efficace pour extraire les **mots** (= oligonucléotides entre 5-10pb avec un certain niveau de substitutions accepté à certaines positions)

exceptionnels dans les séquences nucléotidiques. Son principe se résume en quatre principales étapes qui sont similaires dans les autres algorithmes.

a- Pour chaque mot de taille définie (ex : hexanucléotide, k=6) est estimé une **probabilité *a priori*** (= la probabilité de trouver un mot en particulier dans une position donnée) à partir de la fréquence observée du même mot dans un jeu de contrôle. Celui-ci est, par défaut, un modèle de Markov dont l'ordre de transition des nucléotides est estimé à partir des séquences elles-mêmes. Il est également possible d'utiliser un jeu de contrôle personnalisé.

b- La p-value du mot (= la probabilité d'observer au moins x occurrences de ce mot) est ensuite calculée par une loi binomiale. Elle permet de donner une estimation du risque de faux positif = **FP** (risque de considérer un mot comme significatif alors qu'il ne l'est pas).

c- Plusieurs milliers de tests sont alors effectués pour déterminer les mots surreprésentés. Si le seuil de la p-value n'est pas assez restrictif, le risque d'accepter des FP est plus élevé. Pour corriger ce problème, il est nécessaire de calculer la **e-value** (= nombre attendu de FP correspondant à cette p-value).

e-value = p-value * nombre de mots

d- Au final, le Sig calculé par la transformation en log de la e-value **: Sig = -$\log_{10}$(e-value).**

Les mots découverts sont classés selon leurs scores en p-value, e-value et Sig. Les premiers mots sont les meilleurs trouvés par l'algorithme et vont servir de graine pour construire une description probable du motif. Ils sont assemblés (outil '*pattern-assembly*' : http://rsat.ulb.ac.be/pattern-assembly) puis convertis en une matrice poids position (outil '*convert matrice*' : http://rsat.ulb.ac.be/convert-matrix) pour indiquer la variabilité des résidus à chaque position du motif.

2) Dyad-analysis (van Helden *et al.* 2000b)

Il a été développé pour des études spécifiques de CHIP-Seq concernant la recherche de motifs de fixation de FT. En effet, certains FT dimériques reconnaissent des **dyads**, c'est-à-dire des paires de petits oligonucléotides (3-4pb) séparés par un espace de largeur fixe mais de contenu variable (par exemple : CTAn{10}TGG). Le principe du dyad-analysis est le même que celui de l'oligo-analysis sauf qu'il détecte les dyads surreprésentés dans un jeu de séquences.

3) Position-analysis (van Helden *et al.* 2000a)

Ce programme permet de calculer la distribution positionnelle des oligonucléotides

dans le jeu de séquence et de repérer lesquels s'éloignent significativement d'une distribution homogène. Il peut être utile pour détecter les motifs avec un biais positionnel dans de larges jeux de données (par exemple, plus de mille séquences par exemple). Nous nous sommes rendu compte au cours du stage que cet algorithme n'était pas adapté pour des données CLIP-Seq car il prend en référence le milieu des pics fusionnés, ce qui propre aux analyses CHIP-Seq.

Une fois les motifs obtenus, Peak-motifs va les comparer avec des bases de données publiques (**bd**) contenant des motifs de fixation associés avec des FTs connus. Pour notre analyse, la base de données JASPAR associée aux vertébrés (Mathelier *et al.* 2013) a été choisie.

Résultats

L'outil « Peak-motifs » de RSAT nous a permis d'obtenir une description statistique et nucléotidique des jeux d'entrée (**Figure 3** et **4**), ainsi que de découvrir les principaux motifs présents dans les jeux de données (**Tableau 1, 2** et **3**). Les résultats de l'oligo et du dyad-analysis obtenus entre les différents jeux ont été comparés afin de déterminer les meilleures associations de jeux dans la découverte de motif. Ainsi, trois comparaisons ont été réalisées :

- Jeux positifs sans jeux négatifs.
- Jeux positifs Pics_publi et Pics_new_all testés contre les quatre jeux négatifs.
- Jeux positifs Pics_new_monomères et Pics_new_dimères testés contre les quatre jeux négatifs.

1) Analyse et comparaison des jeux positifs

- **Nombre et longueur des pics**

Nous constatons tout d'abord que le **nombre** de pics du jeu pics_publi par rapport aux nouveaux jeux de données positifs générés à partir des données brutes de séquençage (pics_new_monomère, pics_new_dimère et pics_new_all) est totalement différent (environ 50000 contre quelques milliers) (**Figure 3**). Cette différence peut être dû au fait que le filtre de qualité utilisé (qualité des reads >20) a eu comme incidence d'éliminer une grande partie des reads.

Cependant si les pics sont moins nombreux dans les nouveaux jeux positifs, ils sont **plus longs :** en moyenne 191 à 567 nucléotides contre 27 nucléotides. Cette augmentation de la taille des pics est plutôt rassurante sachant que d'après les analyses structurales du site de liaison de PTBP1 au moins 30 nucléotides sont nécessaire à la fixation de la protéine. Par ailleurs, la taille des pics correspondant à la protéine PTBP1 fixée sur l'ARN en tant que

dimère est trois fois supérieure à ceux de la protéine PTBP1 monomère (566 nucléotides contre 191), ce qui semble cohérent d'un point de vue biologique.

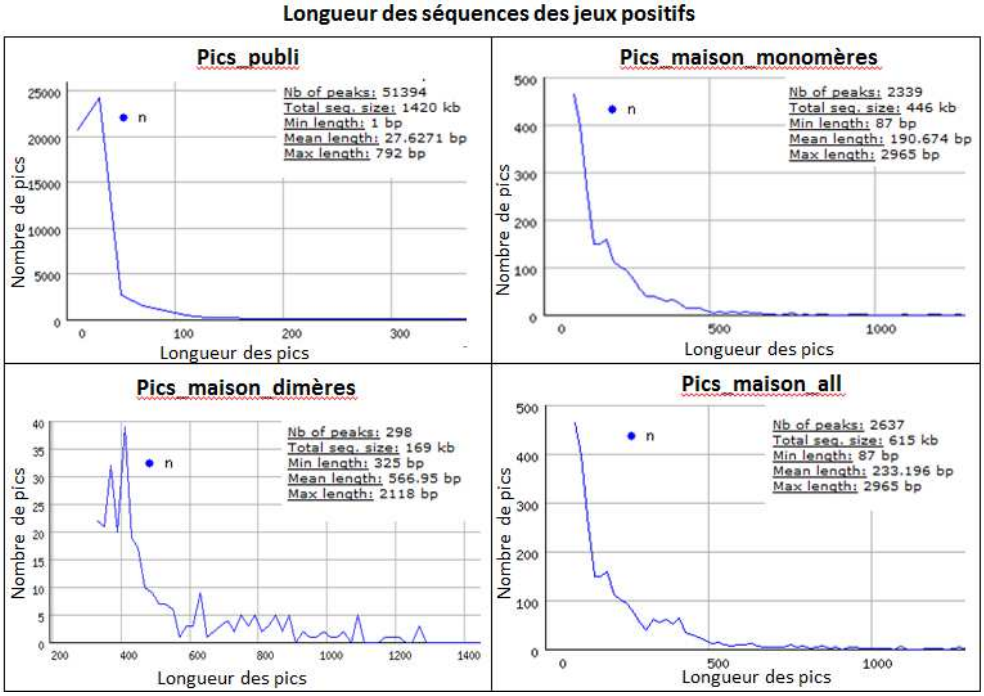


Figure 3 : Statistiques descriptives des quatre jeux positifs.

- **Composition nucléotidiques des pics**

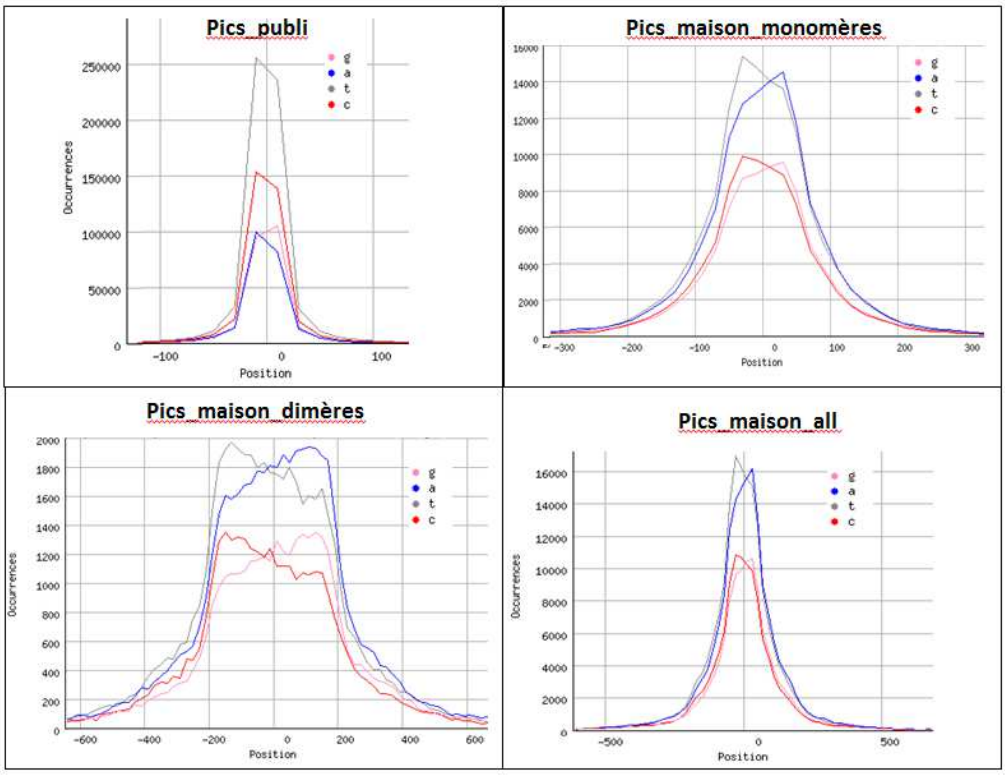


Figure 4 : Composition nucléotidique des quatre jeux positifs.

Nous constatons que Pics_publi possède en majorité des résidus T et C alors que les nouveaux jeux positifs ont des compositions plus équilibrées avec une majorité de T et de A (**Figure 4**). Comme les reads de départ sont les mêmes, cette différence de composition est due aux différentes méthodes de peak-calling. Pour les données de pic_publi aucune information n'est donnée sur la méthodologie employée.

- **Motifs présents dans les jeux positifs analysés sans jeu de contrôle négatifs**

L'algorithme « oligo-analysis » sur les jeux positifs donne des scores de significativité médiocre (Sig<10) chez tous les jeux positifs et des motifs peu ressemblant à ceux proposés dans la littérature.

En revanche, l'algorithme « dyad-analysis » donne des scores de significativité très élevés notamment à partir du jeu de données pic_publi. Globalement, pour les deux algorithmes, il donc apparaît utile d'utiliser des jeux de contrôle négatifs pour augmenter la significativité des motifs produits.

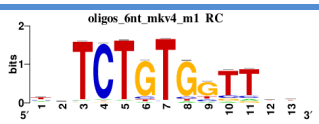
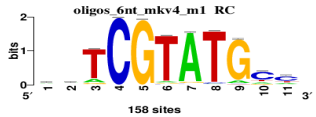
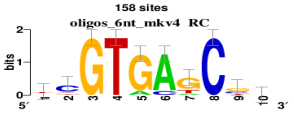
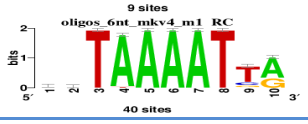
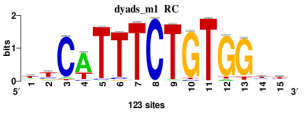
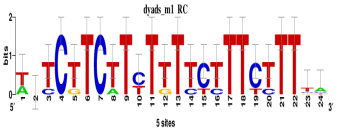
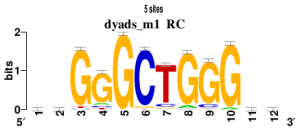
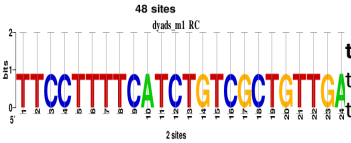
Analyse	Jeu positif	1er Motif	3 premiers mots/dyads	Sig	Match bd	Test Pearson
Oligo	Pics_publi		**tctgtg** cttatt ttctca	9.3 6.3 6.3	RUNX1 RUNX2	0.75 0.73
	Pics_new_monomères		**cgtatg** taaaat taattt	4.8 3.3 1.7	No match	.
	Pics_new_dimères		**gtgagc**	1.1	No match	.
	Pics_new_all		**taaaat** cgtatg taattt	3.9 3.4 3.0	Prrx2 Pdx1 ARID3A	0.48 0.41 0.49
Dyad	Pics_publi		**ttctgt** tttctg tttcct	350.0 350.0 350.0	RUNX1 SPIB REL RELA	0.53 0.44 0.43 0.41
	Pics_new_monomères		**tttctt** ttctct tctgtc	104.6 95.3 78.4	No match	.
	Pics_new_dimères		**gctggg** ctcn{4}tcc ttttct	**23.3** **19.8** **19.6**	SP1 Foxa2	0.65 0.52
	Pics_new_all		**tttctt** ttctct tctgtc	121.2 111.5 90.8	No match	.

Tableau 1 : Jeux positifs sans jeux négatifs contrôle testés avec oligo-analysis et dyad-analysis.

2) Analyse des jeux négatifs

Afin de vérifier que les jeux de données négatifs n'ont pas de biais de composition, Peak-motif a été testé sur chacun de ces jeux (cf **Annexes**). Comme aucun motif n'apparait, ils sont donc considérés comme statistiquement valides (neutres).

3) Analyse des jeux positifs à l'aide de jeux de contrôle négatifs

Il s'agit ici de tester l'influence des différents jeux de contrôle négatifs. Seuls les jeux pics_publi et pics_new_all ont été utilisés afin de pouvoir comparer les scores et les motifs obtenus. La première évidence est que les scores de significativité ont considérablement augmenté avec l'utilisation de jeux contrôle négatifs.

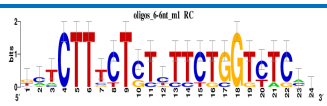
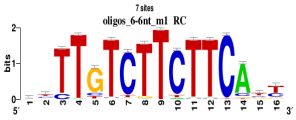
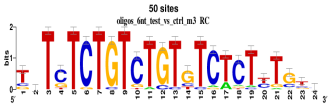
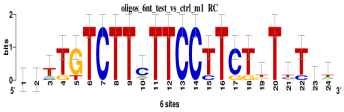
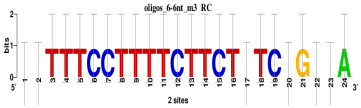
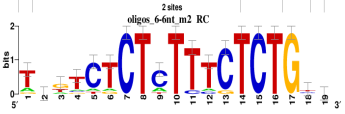
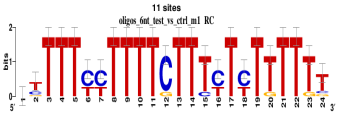
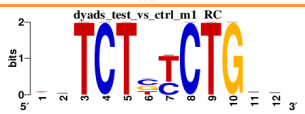
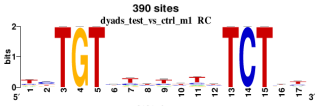
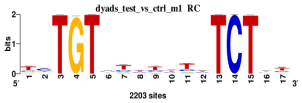
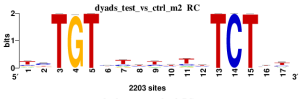
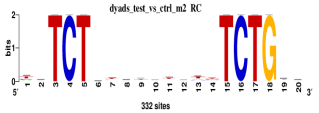
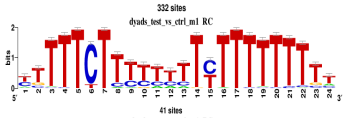
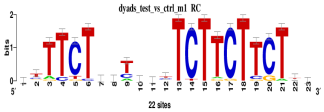
Avec le jeu pic_publi, dans les 2 analyses, les scores de significativité sont au maximum quelque soient les jeux négatifs sauf pour le jeu négatif shuffle qui a des scores légèrement moins bons.

Avec le jeu pics_new_all, les scores sont moins bons dans les deux analyses sauf avec le jeu négatif random_sequence qui donne un motif dyad avec une significativité maximale. Le choix du jeu négatif a donc un impact fort sur la spécificité du motif.

Si l'on compare les motifs fréquents trouvés, les mots trouvés en fonctions des jeux positifs ne sont pas les mêmes sur les deux jeux. Ils sont globalement plus diversifiés avec pics_publi alors qu'avec les nouveaux jeux, on trouve fortement le dyad tct{n}tct et des mots de type ttctct conforme au motif UYUYU (tytyt) de la littérature.

Les dyads trouvés avec pic_publi sont globalement de type ttc{n}ttc, alors que ceux trouvés avec pics_new_all sont majoritairement de type tct{n}tct. Par rapport à ceux de la littérature, les motifs trouvés avec l'algorithme oligo-analysis pour pics_new_all semblent mieux correspondre à une alternance de tytyt que ceux de pics_publi.

Au vue de cette première analyse, il n'apparait pas qu'un jeu de contrôle négatif soit plus pertinent que les autres.

Analyse	Jeu positif	Jeu négatif	1er Motif	3 premiers mots/dyads	sig	Match bd	Test Pearson
Oligo	Pics_publi	Shuffleseq		**atctgt** atctgt cttcag	319.0 309.0 309.0	No match	.
		Random_ sequence		**atctct** tttctg ttgtct	350.0 350.0 350.0	Sox3 Sox6	0.41 0.40
		Random_ genome_ fragments		**ctccct** tttctg tccttc	350.0 350.0 350.0	No match	.
		Hela_genic_ fragments		**tttctg** ttgtct tccttc	350.0 350.0 350.0	Sox3 Sox6	0.43 0.41
	Pics_new_all	Shuffleseq	Pas de motif	**tctctg** ttctct tttctg	283.2 195.6 186.5	No match	.
		Random_ sequence		**tttctt** tcttct ttttct	208.4 190.3 183.8	STAT1 STAT2	0.42 0.40
		Random_ genome_ fragments		**ttctct** ctttct tctttc	114.0 99.6 94.5	No match	.
		Hela_genic_ fragments		**ttctct** ctttct tcttct	183.8 174.5 169.6	No match	.
Dyad	Pics_publi	Shuffleseq		**tctn{1}ctg** tctn{10}ctg tctn{4}ttc	350.0 350.0 350.0	No match	.
		Random_ sequence		**ttcn{2}ttc** ttcn{13}ctt tctn{10}ttg	350.0 350.0 350.0	No match	.
		Random_ genome_ fragments		**ttcn{2}ttc** cttn{7}ctg gagn{4}agg	350.0 350.0 350.0	No match	.
		Hela_genic_ fragments		**ttcn{2}ttc** ttcn{13}ctt ttcn{17}tct	350.0 350.0 350.0	Gata3 Gata1 Gata4	0.42 0.56 0.48
	Pics_new_all	Shuffleseq		**tctn{1}tct** tctn{2}ctg tctn{3}tct	285.9 245.6 209.5	No match	.
		Random_ sequence		**tctn{6}tct** tctn{9}tct tctn{3}tct	350.0 350.0 350.0	No match	.
		Random_ genome_ fragments		**tctn{6}tct** tctn{9}tct tctn{1}tct	175.8 156.1 151.8	No match	.

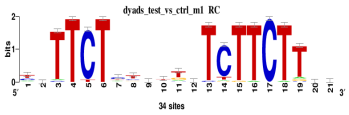
Hela_genic_ fragments		**tctn{6}tct** tctn{1}tct tctn{9}tct	**219.2** **218.7** **217.4**	No match

Tableau 2 : Jeux positifs testés avec avec des jeux de contrôle négatifs testés avec oligo-analysis et dyad-analysis

4) Pertinence de l'affinement des jeux négatifs

L'un des autres objectifs principaux du stage a été de tester différents jeux négatifs pour vérifier s'ils possédaient une influence sur la pertinence des motifs trouvés. Lorsque nous testons un jeu positif sans contrôle, nous ne retrouvons quasiment pas de mots ou de dyads significatifs. En revanche, avec un jeu de contrôle négatif, il nous est possible de cibler plus efficacement les régions pyrimidiques du jeu testé. Ainsi, l'utilisation d'un jeu négatif apparait primordiale pour permettre aux algorithmes d'identifier des éléments surreprésentés non dus au hasard.

Au total, nous avions une première catégorie de jeux négatifs aléatoires sans réel « sens biologique » en particulier (Shuffleseq et Random_sequence) et une deuxième catégorie qui utilise des séquences biologiques réelles pour la construction des jeux (Random_genome_fragments et Hela_genic_fragments). Au regard des résultats, aucune des deux catégories ne semble significativement meilleure que l'autre. D'un point de vue méthodologique, on retiendra donc qu'il est important d'utiliser un jeu de contrôle mais que le choix de ce jeu n'est pas très sensible. D'un point de vue biologique, on peut remarquer que l'analyse des pics_new fait ressortir les dyads.

5) Comparaison des jeux monomères et dimères

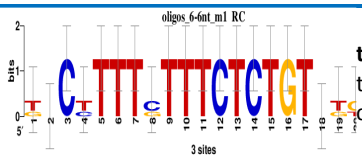
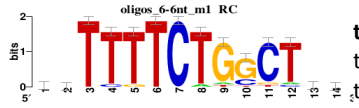
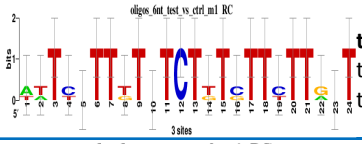
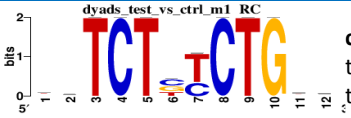
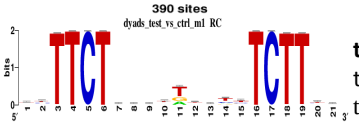
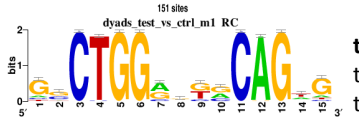
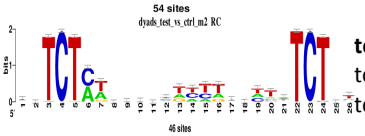
Analyse	Jeu positif	Jeu négatif	1er Motif	3 premiers mots/dyads	Sig	Match bd	Test Pearson
Oligo	Pics_new monomères	Shuffleseq		**tctctg** ttctct ctttct	**238.3** **181.8** **132.1**	Gata1	0.42
		Hela_genic_fragments	Pas de motif	**ttctct** ctttct tcttct	**167.8** **188.4** **157.8**	No match	
	Pics_new dimères	Shuffleseq		**ttttct** tttcct ttctct	**242.6** **153.2** **151.2**	Klf4	0.45
		Hela_genic_fragments		**tctctt** tcttct ttttct	**28.4** **26.7** **24.2**	No match	
Dyad	Pics_new monomères	Shuffleseq		**ctgn{2}tct** tctn{1}tct tctn{2}ctg	**178.4** **173.1** **165.0**	No match	
		Hela_genic_fragments		**tctn{9}tct** tctn{1}tct tctn{5}tct	**213.3** **201.2** **199.6**	No match	
	Pics_new dimères	Shuffleseq		**tggn{4}cag** tctn{1}tct tctn{3}tct	**61.0** **60.5** **49.7**	Hand1 Gata3	0.44 0.52
		Hela_genic_fragments		**tctn{16}tct** tctn{6}tct tctn{5}ttc	**49.83** **36.48** **35.35**	No match	

Tableau 3 : Jeux positifs Pics_new_ monomères et Pics_new_dimères testés avec jeux de contrôle négatifs par Oligo et Dyad-analysis

On observe peu de motifs significatifs pour les dimères, l'effectif de 298 pics étant certainement trop faible pour avoir des résultats pertinents. La comparaison entre les deux jeux monomères et dimères n'est donc pas vraiment réalisable.

Discussion

Du point de vue de notre objectif biologique qui était d'affiner les connaissances des motifs de fixation de la protéine PTBP1 à l'ARN, les motifs identifiés ressemblent globalement à ceux déjà identifiés dans la littérature. Une nouveauté intéressante est la récurrence du motif dyad mise en évidence sur le jeu pics_new qui pourrait correspondre à la fixation des RRMs de la protéine.

tctn{1}tct **tctn{3}tct** **tctn{6}tct** **tctn{9}tct**

Du point de vue de notre objectif méthodologique qui était de tester l'adéquation de RSAT aux données CLIP-Seq, il apparaît que la suite RSAT est globalement bien adaptée à la découverte de motifs au sein de données CLIP-Seq. Le pipeline Peak-motifs possède de nombreux avantages pour des recherches exploratoires en CLIP-Seq : il est facile d'utilisation, rapide, accepte de grands jeux de données en entrée et repose sur des tests statistiques bien adaptés à l'exploration de données biologiques. Sa particularité d'être modulaire permet de contrôler chaque étape de l'analyse. Cependant, il semble moins adapté à des jeux de données de faible effectif (cf le peu de résultats avec le jeu de données pics_new_dimères).

Les jeux négatifs proposés par RSAT correspondant aux contraintes des données de CHIP-Seq, nous avons choisis de construire des jeux de données spécifiques au CLIP-Seq cependant ces jeux n'ont pas semblé plus performants dans le cadre de cette étude.

Un point important à prendre en compte dans cette étude est la qualité des données de séquençage initiales. La ré-analyse des données CLIP-Seq de 2009 avec les critères de qualité actuels a pour conséquence la perte d'une grande partie des reads (99% de perte globale) ce qui entraine une perte de profondeur d'information qui nuit considérablement à la détection correcte des motifs. Cette étude pourrait donc certainement être reconduite avec profit sur des données de séquençage plus récentes afin d'augmenter la longueur et la qualité des reads.

Conclusion

Notre étude exploratoire sur la fixation de la protéine PTBP1 à l'ARN a validé l'outil RSAT comme un outil adapté au traitement des données CLIP-seq. Cette étude nous a permis de créer plusieurs jeux positifs et négatifs et de les tester les uns contre les autres. Il a ainsi été possible de retrouver des motifs presque similaires à ceux trouvés dans la littérature et d'en dégager de nouveaux. Les nouveaux motifs surreprésentés mis en évidence sont de type dyads (CTC...CTC). Nous n'avons cependant pas réussi à identifier des motifs spécifiques entre les jeux de données monomères et dimères, certainement par manque de données. Une des perspectives de cette étude est la recherche des motifs trouvés sur des séquences cibles de PTBP1 afin de renforcer leur crédibilité, la finalité étant d'aboutir à la prédiction in silico de la fixation de la PTBP1 à partir de génomes entiers annotés. Une autre perspective est de continuer la démarche exploratoire des méthodes d'analyse des CLIP-seq sur d'autres protéines PTB et d'autres jeux de données..

Annexes

Présentation de la structure d'accueil Inria/Irisa

Le centre de recherche publique INRIA (Institut National de Recherche en Informatique et Automatique, http ://www.inria.fr) est un institut français spécialisé en mathématiques et informatique. L'INRIA possède des travaux de recherches appliquées au monde de l'entreprise et réalise souvent des projets en partenariat avec des industries de toutes tailles ou des organismes particuliers (hôpitaux, centres de recherche, etc.).

L'INRIA de Rennes est associé avec l'unité IRISA (Institut de Recherche en Informatique et Systèmes Aléatoires, UMR 6074) spécialisée dans la recherche en informatique, en traitement du signal et des images, et en robotique. Ses plateformes technologiques d'expérimentation servent de supports à la recherche dans des domaines phares tels que la bio-informatique, l'imagerie médicale, le biomédical et la robotique. L'IRISA comprend 750 personnes (des chercheurs du CNRS, de l'Inria, de l'université de Rennes 1 et des personnes dédiées au support et à l'accompagnement de la recherche) et 38 équipes de recherche.

J'ai réalisé mon stage au sein de l'une de ces équipes, DYLISS (Dynamics, Logics and Inference for biological Systems and Sequences, http://www.irisa.fr/dyliss) dont la responsable d'équipe est Anne Siegel. L'axe de recherche principal de DYLISS est de caractériser les acteurs génétiques d'espèces non modèles qui contrôlent la réponse phénotypique confronté à leur environnement. Mon stage a été encadré par Catherine Belleannée dont les recherches portent sur l'utilisation de modèles grammaticaux pour modéliser les séquences biologiques (au niveau structural et séquentiel).

Bilan personnel du stage

Durant mon stage, j'ai appris à collaborer avec des chercheurs appartenant à plusieurs équipes qui possèdent des thématiques de recherche variées. J'ai beaucoup apprécié d'intégrer une équipe de recherche aussi conviviale que DYLISS et j'ai été très intéressée par les séminaires proposés par les équipes de l'Irisa. Mon encadrante, Catherine Belleannée, a aussi été très présente, très cordiale, et m'a donné de nombreux conseils pour m'organiser efficacement dans mon travail, ce qui m'a beaucoup apporté. Je me suis rendue compte qu'en recherche (en bio-informatique plus particulièrement), les données peuvent s'accumuler rapidement et donc il m'a fallu une bonne organisation dans la gestion de ces données.

Aborder de manière pratique les notions théoriques vues durant mes enseignements de Master 1 est pour moi la meilleure façon de les assimiler. J'ai pu ainsi acquérir des connaissances supplémentaires sur les motifs biologiques et sur leurs techniques de détection. J'ai également amélioré mes notions sur le travail en environnement Linux et sur le langage de programmation Python.

Pour conclure, je suis très satisfaite de ce stage qui m'aura permis de prendre un peu plus confiance en moi de manière générale et qui a aussi confirmé ma volonté de travailler dans le milieu de la recherche plus tard.

Comparaison des jeux négatifs seuls

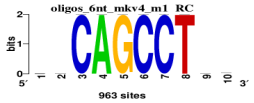
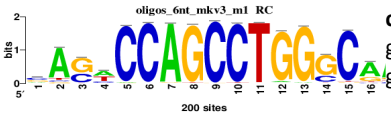
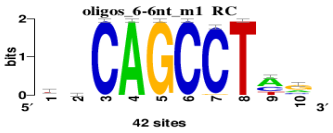
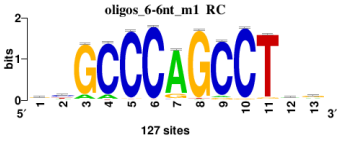
Jeu négatif	1er motif	3 premiers mots	sig	Nb mots signif	Match bd	Test Pearson
Shuffleseq	Not a single significant pattern					
Random_ sequence		**cagcct** gtaatc tgcagt	**46.5** **41.1** **41.0**	80	Mafb	0.44
Random genome fragments	Pas de motif	**gttttc** gacttg	2.7 0.5	0		
Hela_genic fragments monomères		**cagcct** gggagg gcagtg	**26.2** **22.5** **22.1**	7	FOXI1 Foxa2 FOXL1 MZF1_1-4	0.62 0.60 0.45 0.45
Hela_genic fragments dimères		**cagcct** ttgaga gaaaag	5.8 2.3 2.2	0	MEF2C MEF2A Crx STAT1 IRF1 Mafb	0.73 0.72 0.53 0.46 0.43 0.41
Hela_genic fragments _all		**cagcct** gggagg gcagtg	**38.5** **30.8** **30.0**	16	SP1 KLF5 ZEB1 MEF2C MEF2A MZF1_1-4	0.71 0.78 0.52 0.58 0.42 0.45

Tableau 4 : Jeux négatifs testés sans jeux de contrôle avec oligo-analysis

Bibliographie

Amir-Ahmady B., Boutz P.L., Markovtsov V., Phillips M.L. & Black D.L. (2005). Exon Repression by Polypyrimidine Tract Binding Protein. *RNA*, 11, 699-716.

Defrance, M., R. Janky, et al. (2008). "Using RSAT Oligo-Analysis and Dyad-Analysis Tools to Discover Regulatory Signals in Nucleic Sequences." Nat Protoc **3**: 1589-1603.

Fujita P.A. et *al.* (2010). The UCSC Genome Browser database: update 2011. *Nucl. Acids Res.*, 10, 1-7.

Garcia-Blanco M.A., Jamison S.F. & Sharp P.A. (1989). Identification and Purification of a 62,000-Dalton Protein that Binds Specifically to the Polypyrimidine Tract of Introns *Genes & Development*, 3, 1874-1886.

Glisovic T., Bachorik J.L., Yong J. & Dreyfuss G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett*, 18, 1977-1986.

Han A., Stoilov P., Linares A.J., Zhou Y., Fu Z.-D. & Black D.L. (2014). De Novo Prediction of PTBP1 Binding and Splicing Targets Reveals Unexpected Features of Its RNA Recognition and Function. *PLOS Computational Biology*, 10, 1-18.

Mathelier A.,*et al.* (2013). JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucl. Acids Res.*, 10, 1-6.

Oberstrass F.C., Auweter S.D., Erat M., Hargous Y., Henning A., Wenter P., Reymond L., Amir-Ahmady B., Pitsch S., Black D.L. & Allain F.H.-T. (2005). Structure of PTB Bound to RNA: Specific Binding and Implications for Splicing Regulation. *Science*, 309, 2054-2057.

Sawicka K., Bushell M., Spriggs K.A. & Willis A.E. (2008). Polypyrimidine-tract-binding protein: a multifunctional RNA-binding protein. *Biochem. Soc. Trans.*, 36, 641-647.

Ule J., Jensen K., Mele A. & Darnell R.B. (2005). CLIP: A Method for Identifying Protein-RNA Interaction Sites in Living Cells. *Methods*, 37, 376-386.

van Helden J. (2003). Regulatory Sequence Analysis Tools. *Nucl. Acids Res.*, 31, 3593-3596.

van Helden J., André B. & Collado-Vides J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol*, 281, 827-842.

van Helden J., del Olmo M. & Pérez-Ortín J.E. (2000a). Statistical Analysis of Yeast Genomic Downstream Sequences Reveals Putative Polyadenylation Signals. *Nucl. Acids Res.*, 28, 1000-1010.

van Helden J., Rios A.F. & Collado-Vides J. (2000b). Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucl. Acids Res.*, 28, 1808-1818.

Xue Y., *et al.* (2009). Genome-wide Analysis of PTB-RNA Interactions Reveals a Strategy Used by the General Splicing Repressor to Modulate Exon Inclusion or Skipping. *Molecular Cell*, 36, 996–1006.

Résumé

La PTBP1 est une protéine de liaison à l'ARN qui participe activement dans la régulation de l'épissage alternatif. Elle contribue au développement de l'ARN mature en formant des complexes avec l'ARN pré-messager à l'aide de quatre motifs de liaison, appelés RRMs. L'identification des sites de fixation de la PTBP1 passe par la méthode du CLIP-Seq qui va détecter les séquences ARNs liées à la protéine. Ensuite, une analyse de découverte permet de caractériser les signaux biologiques présents dans le jeu de séquences CLIP-Seq. L'objectif biologique du stage concerne le traitement des données publiques de CLIP-Seq de PTBP1 en utilisant de méthodes d'analyses récentes afin d'affiner la reconnaissance des motifs de fixation de PTBP1. La finalité est d'améliorer la prédiction *in silico* de la fixation de la PTBP1 sur des ARNs. Un second objectif méthodologique est d'explorer l'outil «Peak-motifs» de la suite logicielle RSAT, dédié à la découverte de motifs, sur des données CLIP-Seq de PTBP1. Cette analyse a conduit à la construction de plusieurs jeux de données positifs et négatifs afin d'affiner la recherche de motifs. Nos résultats démontrent la compatibilité de RSAT, initialement conçu pour l'analyse CHIP-Seq, aux données CLIP-Seq. L'amélioration des jeux positifs et négatifs nous a aidé à discriminer de nombreux éléments surreprésentés dans les analyses. Des expériences supplémentaires à partir de gènes cibles connus de la PTBP1 viseront à valider les motifs identifiés au cours du stage.

Mots-clés : CLIP-Seq, Découverte de motif, RSAT, Peak-motifs, PTBP1, protéine de liaison à l'ARN

Abstract

The PTBP1 is a RNA binding protein actively involved in the alternative splicing regulation. It contributes to the development of mature RNA by forming complexes with the pre-messenger RNA via four RNA recognition motifs (RRM). A first step to identify the PTBP1 binding sites is use CLIP-Seq method that detects RNA sequences associated with the protein. In a second step, a pattern discovery analysis can characterize biological signals present in the set of sequences CLIP-Seq. The objective of the study is in the treatment of public data CLIP-Seq of PTBP1 by using recent analytical methods in order to refine knowledge about PTBP1 fixation patterns. The aim is to improve the *in silico* prediction of RNA-binding of PBTP1. A second objective is to explore the "Peak-motifs" tool of the RSAT suite dedicated to the discovery of patterns on PTBP1 CLIP-Seq data. This analysis led to the construction of several positive and negative sets of data to refine the search patterns. Our results demonstrate the compatibility of RSAT (originally designed for the analysis CHIP-Seq) to the CLIP-Seq data. The improvement of positive and negative sets helps us to discriminate many elements overrepresented in analyzes. Additional experiments with a list of known PTBP1 target genes will aim to validate the patterns identified during the internship.

Keywords: CLIP-Seq, Pattern discovery, RSAT, Peak-motifs, PTBP1, RNA binding protein